数理解析研究所講究録
第1853巻 2013年 124-133

分子進化実験からの適応度地形情報の抽出

埼玉大学　理工学研究科　相田拓洋
Takuyo Aita
Graduate School of Science and Engineering, Saitama University

1　生体高分子の配列空間と適応度地形について

進化分子工学では、配列の進化の過程を数理的に記述するために、進化を”適応度地形”という概念上の山の”登山（hill climbing）”に喩える。例として、長さnのDNA配列を考える。すべての可能な配列の数は4^nであるが、これらを包括的に扱うための概念上の空間を配列空間 (sequence space) または遺伝子型空間 (genotype space) と呼ぶ。配列上の各部位が配列空間の各座標軸に対応しており（この場合は空間の次元数はn）各座標軸上にはA,T,G,Cの4文字が配置される。従って、任意の配列はこの空間内のある点に対応付けられる（図1）。ここで、任意の2つの配列間の距離はハミング距離 (Hamming distance) で記述するのが一般的である。さらに配列空間内の各点xに、対応する適応度$f(x)$をプロットしてできるスカラー場を適応度地形 (fitness landscape) と呼ぶ [1] [2]。すなわち、適応度地形は配列から適応度への関数関係を表現している。配列を歩行者に見立てたとき、配列に生じる突然変異が「配列空間内の移動」を表現し、このとき、変異が生じた部位の数（ハミング距離）が「歩行者の歩幅」に相当する。つまり、ランダム変異と淘汰の繰り返しによるダーウィン進化の過程は「適応度地形上の”杖を持った盲人”の山登り」と表現できる。ここで盲人は杖を使い周囲の高低を探り、より高い方向へ進む。この山登りを適応歩行 (adaptive walk) と呼ぶ（図2）。盲人である理由は「変異は偶然の揺らぎに起因するものであり、配列自体は進むべき方向を知り得ない」からであり、”杖を持っている”理由は「ランダム変異という”杖”を使って、周囲の高低を探る」からである。すなわち、進化にとって適応度地形の形状は”羅針盤”の役目を果たしていると言える。

実際、適応度地形の構造は歩行者である配列（DNA）の表現型分子（タンパク質）とその周囲の物理化学的環境（ターゲット分子や溶媒の物理化学的特性・条件）に依存し、第1近似では一意的に決まると考えてよいだろう。しかし厳密には、環境の揺らぎや同一配列における表現型多型がある場合には、一意的には決まらない場合もある。すなわち、適応度地形の構造は、実験条件を与えるとほぼ決まると考えてよいが、実験者の意のままに制御することは難しく、あくまで物理化学法則により与えられる構造と認識すべきであろう。ゆえに、適応度地形は生体高分子の”進化的物性（進化的属性：evolutional attribute）”といえる。

2 適応度地形の探査法とその適用例

適応歩行を効率良く行うには、対象となる適応度地形の起伏の度合いなどの統計的性質を知り、その情報に基づいた進化戦略を立てるのが望ましい。そこで、天然配列の周囲の探査法と、山麓からのより広範な領域の探査法を紹介する。

(1) 天然配列の周囲（山頂付近）の探査法： 基本的には突然変異の効果の相加性の程度を調べることが重要である。最も単純なアプローチは、変異効果の相加性に基づいた富士山型地形モデルを仮定する方法である。これは、ある配列"$\mathbf{x} \equiv \mathrm{x_1 x_2 \cdots x_n}$"の適応度関数を、個々の部位の寄与$w_j(\mathrm{x}_j)$の和

$$W(\mathbf{x}) = W_0 + \sum_{j=1}^{n} w_j(\mathrm{x}_j) \tag{1}$$

と仮定して、実験で測定された適応度データにパラメータ適合させる方法である。パラメータ適合後の残差が非相加性の項を表している。より精密な解析法としては、相加的な項に加えて２体以上の相互作用の項を加えたモデルを用いるのが常套手段である [9]。ところで実際には、配列空間の広さに比べたら、実験で測定できる変異体の適応度データの数は極めて少ない。さらに、配列上の全ての部位に渡って置換効果の相加性を調べるのは容易ではない。そこで現実には、重要な複数の部位（例えば、適応度を増加させる部位など）に限定して置換効果の相加性を調べるのが一般的である。すなわち、これらの方法で探査できる領域は天然配列のごく近傍に限られる。（さらに詳しく言うと、複数の部位に限定した探査は「適応度地形を当該座標軸方向に切断した断面」の探査と解釈できる。）

上記の方法で、いくつかの天然タンパク質配列の近傍が探査された。それらの結果からは、これら天然配列の近傍の領域、すなわち、山頂付近は突然変異の効果に概ね相加性が確認された。すなわち、局所的にはスロープが存在することを示唆している。そこで、天然配列の周囲の解析例として、酵素プロリルエンドペプチダーゼを採りあげて解説する [6]。まず始めに、１４箇所の部位で酵素の耐熱性高める置換残基を見出し (具体的には、S19T, E67Q, F70L, S110F, N387K, A388V, S475G, I493V, E496K, N542I, K615R, G652V,S653A, Q656R)、次に、これら１４箇所の置換を複数組み合わせた多重置換体を４５種類作成した。これらの４５種類の配列の適応度を式 (1)（注：この場合、部位の通し番号jはこれら１４箇所に限定して、$n = 14$とする。それら以外の部位は解析には含まれない）でパラメータ適合させたところ、相関係数は０．９３８であった（図3a)。図3aは適応度の実測値とパラメータ適合値の相関を表したものであるが、１つの異常値（◇）が見られる。次に、より真に近いモデルとして適応度関数に２体相互作用項を１つだけ加えること

にした（その可能なペアの数は $14 \times 13/2 = 91$ 種類）。すなわち、式（1）を

$$W(\mathbf{x}) = W_0 + \sum_{j \neq k,l}^{n} w_j(\mathrm{x}_j) + w_b(\mathrm{x}_k, \mathrm{x}_l) \tag{2}$$

と修正した。ここで、最後の項が相互作用項であり、k と l は相互作用する2つの部位を表す。パラメータ適合の結果、「置換残基 E67Q と Q656R が相互作用する」としたモデルの場合に最も適合することが分かり、その場合の相関係数は0．９７１に上がった（図3b)。図3bを見ると、たった1つの相互作用項を加えただけで、異常値（◇）が解消され、全体的に直線に近づいたことが分かる。図3(c)は、式（2）の適応度関数を用いた場合に決定されたパラメータの値を示す。これから、$j = 2$ と $j = 12$ の部位において、野生型残基ペア（"EQ"）のいずれか一方が変異型残基("ER"または"QQ"）になると適応度が大きく増大するが、両方とも変異型残基（"QR"）になると負の相互作用が生じて適応度は激減することが分かる。図3(d)は、この解析で得られた「適応度地形の断面」の最も単純な表現の一つである。最適配列からのハミング距離が大きくなるにつれて配列の適応度が減少する傾向が見られる。すなわちこれは当該地形の断面の概観が富士山型地形に近いことを示している。

当然の流れとして、新たな相互作用項をさらに1つ付け加えることも考えられるが、この場合は「赤池の情報量基準」の観点から不要であることが分かった。（この系における式（2）の適応度モデルの妥当性については文献[6]を参照のこと。）

ところで、部分配列空間内の全ての配列の適応度データがそろっている場合（上記の例では 2^{14} 種類の配列の適応度データが得られた場合）は、よりシステマティックな解析法として、Bahadur展開を利用する方法が有用である[3][4]。任意の配列を $\mathbf{x} = \mathrm{x}_1\mathrm{x}_2\cdots\mathrm{x}_n$ とする。ただし、$\mathrm{x}_j \in \{0,1\}\,(j = 1,2,\cdots n)$ である（例えば、野生型残基を $\mathrm{x}_j = 0$、変異型残基を $\mathrm{x}_j = 1$ とする)。全ての可能な配列の集合を $\mathbf{S}$ $(\equiv \{\mathbf{x}\})$ とする。そこで、次の正規直交関数系 $\psi_i(\mathbf{x})$ $(i = 0,1,2,\cdots,2^n - 1)$ を導入する。まず、$\mathbf{x}$ を変換

$$z_j \equiv 2\mathrm{x}_j - 1 = \begin{cases} -1, & \text{if } \mathrm{x}_j = 0 \\ 1, & \text{if } \mathrm{x}_j = 1 \end{cases}. \tag{3}$$

を用いて1と-1の数値列に変換する。次の正規直交関数系を定義する。

$$\psi_0(\mathbf{x}) = 1 \tag{4}$$

$$(\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \cdots, \psi_n(\mathbf{x})) = (z_1, z_2, \cdots, z_n) \tag{5}$$

$$(\psi_{n+1}(\mathbf{x}), \psi_{n+2}(\mathbf{x}), \cdots, \psi_{n+n(n-1)/2}(\mathbf{x})) = (z_1z_2, z_1z_3, \cdots, z_{n-1}z_n) \tag{6}$$

$$\vdots$$

$$\psi_{2^n-1}(\mathbf{x}) = z_1z_2\cdots z_n \tag{7}$$

この直交関数系には以下の関係が成り立つ。

$$\psi_i(\mathbf{x})^2 = 1 \quad (i = 0, 1, 2, \cdots, 2^n - 1) \tag{8}$$

$$\frac{1}{2^n} \sum_{\mathbf{x} \in \mathbf{S}} \psi_i(\mathbf{x}) \psi_{i'}(\mathbf{x}) = \begin{cases} 1, & \text{if } i = i' \\ 0, & \text{if } i \neq i' \end{cases} \tag{9}$$

$$\frac{1}{2^n} \sum_{i=0}^{2^n-1} \psi_i(\mathbf{x}) \psi_i(\mathbf{x}') = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{x}' \\ 0, & \text{if } \mathbf{x} \neq \mathbf{x}' \end{cases} . \tag{10}$$

この関係を用いて、配列 $\mathbf{x}$ の任意の関数 $f(\mathbf{x})$ は

$$f(\mathbf{x}) = \sum_{i=0}^{2^n-1} w_i \, \psi_i(\mathbf{x}) \tag{11}$$

と展開できる。ここで、w_i は Bahadur 係数であり、$i = 1 \sim n$ の係数は各部位の（相加的な）寄与を、$i = n+1 \sim n + n(n-1)/2$ の係数は各部位ペアの（相互作用の）寄与を与える。全ての配列について $f(\mathbf{x})$ が既知の場合、Bahadur 係数 w_i は

$$w_i = \frac{1}{2^n} \sum_{\mathbf{x} \in \mathbf{S}} f(\mathbf{x}) \, \psi_i(\mathbf{x}) \tag{12}$$

と決定できる。すなわち、$f(\mathbf{x})$ に適応度データ $W(\mathbf{x})$ を代入すれば、各 Bahadur 係数 w_i を求めることができる [4]。

（2）山麓から中腹にかけての広範な領域の探査法： 一方で、より広範囲の領域を探査をする方法の決定打は未だ確立されていない。しかし、前にも述べたＮＫモデルを仮定し、適応歩行の過程を解析することで大まかな描像を得ることは可能かもしれない。ＮＫモデルでは、ある部位における置換の効果が他の K 個の部位に影響を与えるモデルであり、その適応度関数は

$$W(\mathbf{x}) = \sum_{j=1}^{n} w_j(\mathrm{x}_j | \mathrm{x}_{j_1}, \mathrm{x}_{j_2}, \cdots, \mathrm{x}_{j_K}) \tag{13}$$

と与えられる。ここで、$\mathrm{x}_{j_1}, \mathrm{x}_{j_2}, \cdots, \mathrm{x}_{j_K}$ は部位 $j_1, j_2, \cdots j_K$ における残基を表す。すなわち、適応度は関係する $K+1$ 箇所の部位における残基の組み合わせによって決まる。特に $K = 0$ の場合は、前述の富士山型モデルに等しく、K の値が大きくなるにつれて地形表面の凸凹が激しくなる。

適応歩行の手順として最も単純な次の手順を用いることにする。１つの親配列から N 個のランダムな「d 点突然変異体（置換する部位の数が d 個の突然変異体配列）」を作成し、この中で最も適応度の高い変異体を次世代の親配列にする。この歩幅が d の適応歩行を数世代繰り返すとやがて「変異・ランダム抽出・淘汰のバランス」が生じることで定常状態に達する。このＮＫモデル地形上の適応歩行の過程は解析的に数式として求められており、これらの数式を適応歩行データにパラメータ適合することで、相互作用数 K の値などが大まかに推定できる [7]。以下に具体的に説明

する。定常状態での適応度の値は

$$W^* \approx 2\sqrt{\frac{nV \ln N}{d(1+K)}} \tag{14}$$

で与えられる。ここで、V は適応度地形全体に渡る適応度の分散である。一方、平均値は 0 としている。世代 t における歩行者の適応度を W_t と表すと、その期待値は

$$E[W_t] \approx W_0 + \left(1 - \left(1 - \frac{d(1+K)}{n}\right)^t\right)(W^* - W_0) \tag{15}$$

に従う。ここで、既知のパラメータは、配列長 n, サンプリングサイズ N, 置換残基数 d であり、実験で得られる値は、定常値 W^* と適応度の時系列 W_t $(t = 0, 1, 2, \cdots)$ である。そこで、式 (15) を適応度の時系列 W_t $(t = 0, 1, 2, \cdots)$ に適合させることで K の値を推定し、次に式 (14) を用いることで V の値を推定できる。

文献 [5] と [7] では、バクテリオファージの大腸菌への感染能の進化実験について報告している。実験では、バクテリオファージ fd-tet の「パイロットタンパク質 g3p の D2 ドメイン」を 139 アミノ酸からなる 1 つのランダム配列 (RP3-42) に置換して、それを起点として山麓から中腹まで漸進的に進化させた（図 4)。これらの文献では、上記の解析法を適用して、「ファージの感染能の地形」に関する K の値を 21 ～ 27 と推定している。この値が妥当か否かは議論の余地があるが、このかなり大きな値が示唆しているのは「山麓から中腹にかけては突然変異の効果は非相加的であり、地形表面は凸凹が激しい」ということである。

以上の探査結果をまとめると、現実の適応度地形は、「山麓から中腹にかけては凸凹が激しいが、山頂付近ではなだらかなスロープが存在する」と推測できるであろう。この大域的な描像はあくまで仮説であるが、今後の地形探査を行うことで少しずつその真の姿に迫ることができるであろう。

参考文献

[1] J. Maynard-Smith. *Nature* **225**, 563-564.(1970).

[2] C.A. Voigt, *et al. Advances in Protein Chemistry* **55**, 79-160. (2000).

[3] M. Arita M, *et al. Bioinformatics* **18**: S27—S34, (2002)

[4] T. Matsuura, *et al. Molecular System Biology*, **5**: 297 (2009)

[5] Y. Hayashi, *et al. PLoS ONE*, **1**, e96, (2006)

[6] T. Aita, *et al. Biopolymers*, **64**, 95-105, (2002).

[7] T. Aita, *et al. J.theor.Biol.* **246**, 538-550. (2007)

[8] N. Hamamatsu, *et al. Protein engineering design & selection*, **18**, 265-271, (2005)

[9] R. Fox, *et al. Protein Eng.*, **16**, 589-597, (2003)

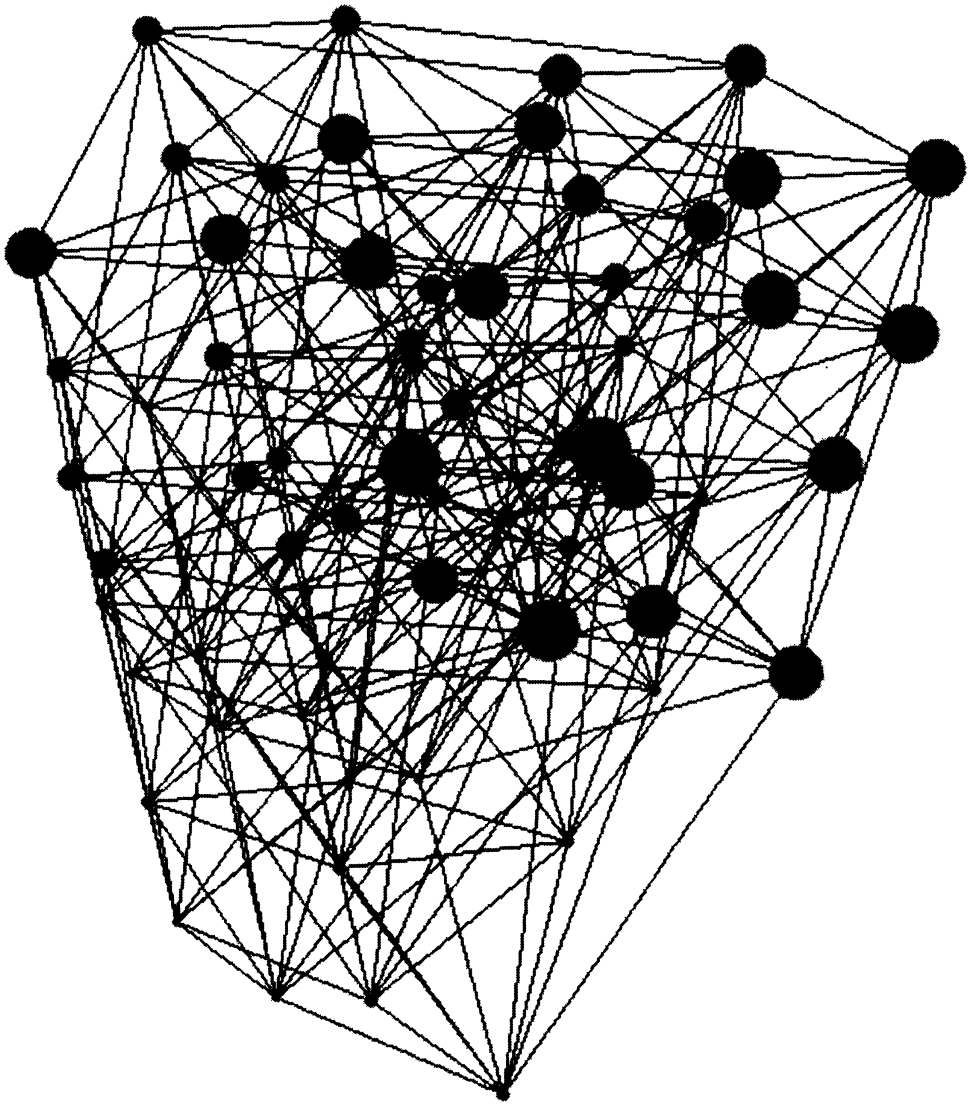

図 1: **配列空間と適応度地形の例。** 文字の種類を λ，配列の長さを n としたとき，全ての可能な配列は λ^n 種類存在する．この各配列をグラフの各頂点に 1 対 1 に対応させ，配列表示で 1 文字だけ異なる頂点間の全てを辺で結ぶ。こうして得られるグラフ構造が配列空間である．配列空間において任意の 2 点間の距離がハミング距離（＝同じ長さの 2 つの配列を比較したときの互いに異なっている文字の数）である．さらに配列空間中の各点 $\mathbf{x}$ に、対応する適応度 $f(\mathbf{x})$ をマップしてできるスカラー場が適応度地形である。ここでは例として、「コドンの配列空間（$\lambda = 4$、$n = 3$)」と「各コドンのコードするアミノ酸の疎水性値を適応度とした適応度地形」を示した。○の大きさが適応度を表す。この地形はほぼ富士山型になっている。

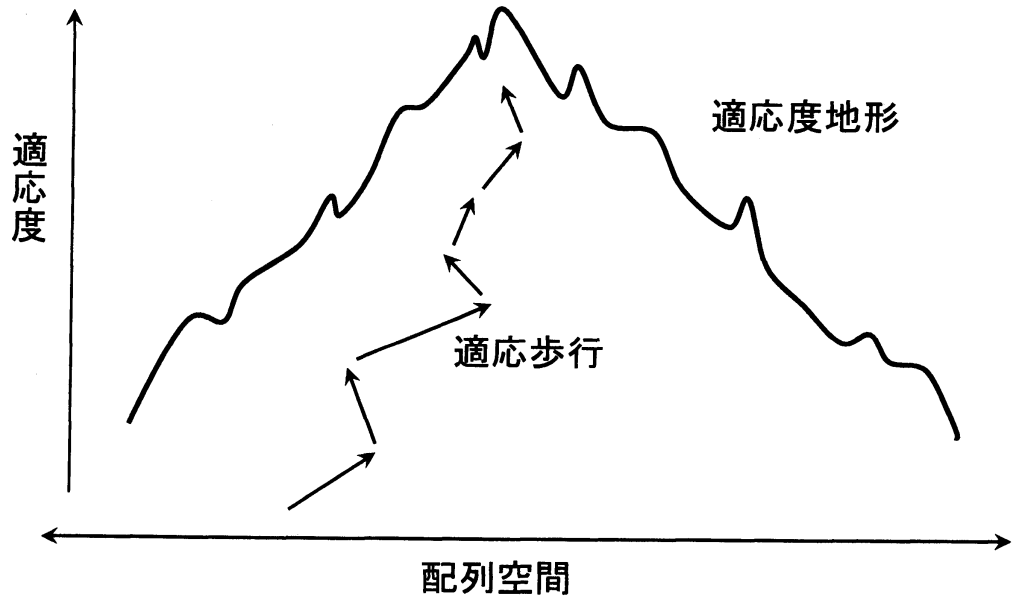

図 2: **適応度地形上の適応歩行。** 1 回の進化サイクルが適応歩行の 1 ステップに相当する。配列の複製において残基置換が生じた部位の数、すなわち、ハミング距離が適応歩行の歩幅に相当する。

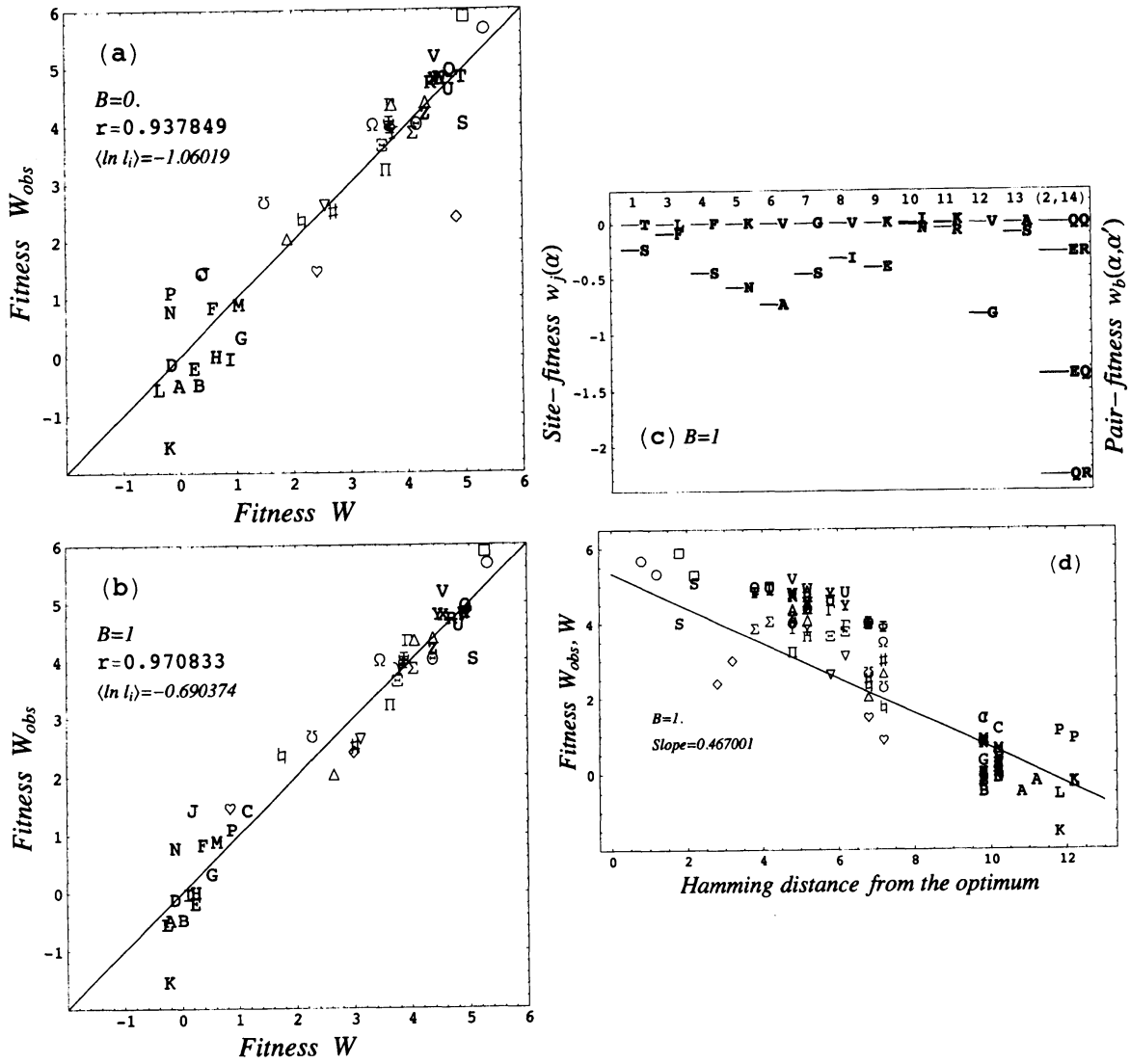

図 3: **プロリルエンドペプチダーゼの耐熱性の地形の解析結果.** (a,b) 縦軸は適応度の実測値 $W_{\rm obs}$、横軸はパラメータ適合で得た適応度の値 W. (a) は式 (1) の適応度関数を用いた場合で、(b) は式 (2) の適応度関数を用いた場合。各シンボルは４５種類の各クローンに対応する. r は相関係数. (c) 式 (2) の適応度関数を用いた場合に決定されたパラメータの値。左から 1〜13 のレーンは部位の寄与 $w_j(\mathrm{x}_j)$ を表し，一番右側のレーンは残基ペアの寄与 $w_b(\mathrm{x}_k, \mathrm{x}_l)$ を表す. 図の上部に示した数字は解析のための部位の通し番号である. アルファベットは一文字表記のアミノ酸を表す. 一番右のレーン (2,14) においては，"EQ"は，部位 2 が E（グルタミン酸）で部位 14 が Q（グルタミン）の意味であり，このペアは野生型残基ペアである. これら 2 つのいずれかが変異型残基（"ER"、"QQ"）になると適応度が大きく増大するが、両方とも変異型残基（"QR"）になると適応度は激減する。(d) 配列空間を 1 次元軸（最適配列からのハミング距離）へ投影することによる適応度地形の表現. 各ハミング距離の位置において，左列の文字は実測値 $W_{\rm obs}$ を，右列の文字はパラメータ適合で得た適応度の値 W を表す. 斜線は地形の平均勾配を表す. (文献 [6] より転載)

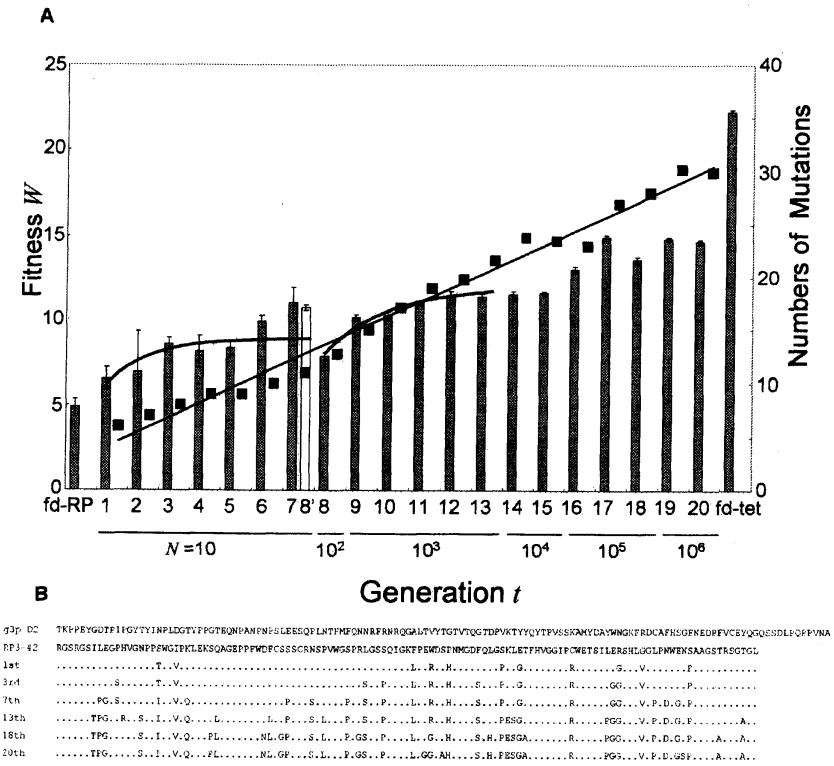

図 4: **バクテリオファージの大腸菌への感染能の進化.** 20 世代に渡る親配列（＝歩行者）の適応度の増加の様子を棒グラフで表した。世代を経るにつれてサンプルサイズ N を 10 から 10^6 まで段階的に増やしていった。一方、■は出発配列 (RP3-42) から蓄積されていった同義置換の数を示す。2 つの指数関数型の曲線は、$N = 10$ と $N = 10^3$ のそれぞれの適応度の時系列に対して式（15）を適合させた結果である。その結果、$K = 21$ が得られた。（文献 [7] では、より精密な解析を行い $K = 27$ と推定した。）（文献 [5] から転載）